基于翻译质量估计的神经网络译文自动后编辑[1]

谭亦鸣 王明文† 李茂西

江西师范大学计算机信息工程学院，南昌 330022；† 通信作者，E-mail: mwwang@jxnu.edu.cn

摘要 译文自动后编辑是利用规则或统计的方法自动纠正机器翻译译文中出现的错误，对提高机器翻译的译文质量有着重要作用。针对译文后编辑中的过度修正问题，在对 WMT 自动后编辑任务训练集语料进行统计和分析之后，我们发现其中超过半数的机器译文仅需少量的编辑操作就可以修改为正确的译文，并提出利用神经网络自动后编辑方法训练专门用于提供少量复合编辑修正和单一编辑类型修正的神经网络后编辑模型。在此基础上，通过建立一个基于翻译质量估计的译文筛选算法将提出的模型与常规的神经网络自动后编辑模型进行联合。在 WMT16 自动后编辑任务测试集上的实验结果表明，与基准系统相比，本文提出的方法显著的提高了机器译文的翻译质量，进一步的实验分析表明了该方法能有效的处理译文过度修正所造成的译文质量下降问题。

关键词 译文自动后编辑；神经机器翻译；机器翻译质量估计；过度修正

中图分类号 TP391

Neural Post-Editing Based on Machine Translation Quality Estimation

TAN Yiming, WANG Mingwen†, LI Maoxi

School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022;

†Corresponding Author, E-mail: mwwang@jxnu.edu.cn

Abstract Automatic post-editing (APE) aims to correct machine translation errors by rule methods or statistical methods; it plays an important role in the application and popularization of machine translation. Through statistical analysis of the training set released by WMT APE Shared Task, we found that more than half of machine translations only need a small amount of edit operations. To reduce over-editing problem, we proposes to make advantage of the neural post-editing (NPE) to build two special models, one is used to provide minor edit operations, the other is used to provide single edit operation, and make advantage of machine translation quality estimation to establish a filtering algorithm to integrate the special models with the regular NPE model into a jointed model. Experimental results on the test set of WMT16 APE shared task show that the proposed approach statistically outperforms the baseline. Deep analysis further confirms that our approach can bring considerable relief from the over-editing problem in APE.

Key words automatic post-editing; neural machine translation; quality estimation of machine translation; overcorrection

机器翻译是利用计算机把一种自然语言自动转换成另一种自然语言的过程[1-3]。近三十年来，在国内外学者的不懈努力下，机器翻译研究取得了飞跃的发展，许多机器翻译方法相继提出，如基于规则的翻译方法、基于实例的翻译方法、传统统计翻译方法和神经机器翻译方法等等。尽管机器翻译方法取得了长足的进步，机器译文的质量也在不断提高，但是机器译文中仍存在少量翻译错误，不经过后编辑它很难在正式文本中得到使用。

译文后编辑方法分为两种，一种是人工后编辑，它的优点是准确，但是人力成本高，时间周期长；另一种是自动后编辑 (Automatic post-editing, APE)，它的优点是方便引入在机器翻译解码时不易于引入的额外知识来提高机器译文的质量，为人工后编辑节省时间。

Knight 和 Chander[4]在采用规则的方法处理日语到英语翻译任务中名词短语前冠词的选择问题时，最早提出自动后编辑的概念。在基于短语的统计机器翻译框架下，Simard 等人[5]提出利用统计后编辑来

[1] 国家自然科学基金(61662031，61462044，61462045)资助

提高机器译文的质量，主要思路是训练一个基于短语的单语统计机器翻译系统把存在错误的机器译文翻译成正确的译文，它利用已标注的人工后编辑译文训练翻译系统，把待编辑的机器译文看做源语言句子，把已经过人工编辑的译文看做目标语言句子构建平行语料。由于他们的工作中没有考虑原翻译知识，Bechara 等人[6]引入源语言句子并把它作为自动后编辑的上下文来构建统计机器翻译系统，为了避免训练数据的稀疏，他们对 GIZA++训练的词汇化概率设置阈值以引入更准确的翻译上下文信息。上面这两种方法分别在不同的语料上证实了它们的效果，为了更一致地对它们性能进行系统性的比较，Chatterjee 等人[8]在大规模的 6 个语言对的语料上采用翻译编辑率 (Translation Edit Rate, TER) [7]打分方法分别对这两种方法进行了对比，发现后者性能稍稍优于前者。

在自动后编辑中一个普遍的问题是过度修正，过度修正是指译文自动编辑的次数超过它实际需要的操作次数，因此，过度修正容易再次引入错误，降低译文的质量。在 WMT15 APE 子任务结果统计中，8 个参加评测系统的后编辑译文质量均低于原始待编辑译文质量[9]，而在 WMT16 APE 子任务结果统计中，12 个参加评测系统中 4 个系统的后编辑译文质量低于原始待编辑译文质量[10]，产生这个现象的一个突出原因是过度修正。

为了减少过度修正，本文通过在 WMT APE 子任务训练集上统计与分析译文后编辑操作次数的分布，以预测测试集中机器译文需要的编辑次数。在此基础上，提出使用最新的神经网络译文后编辑方法对需要少量编辑修正的机器译文建立单独的后编辑模型，并将这个模型与常规的神经网络自动后编辑模型的输出结果组成后编辑 *n*-best 列表，以改进的翻译质量估计方法对译文后编辑 *n*-best 进行打分排序，并选择出最优后编辑译文。我们在 WMT16 APE 测试集上的实验结果表明，排序挑选出的最优后编辑译文质量相对基线系统输出译文的质量有了较大幅度的提高。

1 相关工作

近年来，随着深度学习在自然语言处理中的广泛应用，许多工作将其应用于译文自动后编辑。Pal 等人[11]提出利用双向递归神经网络编码器-解码器模型建立一个单语机器翻译系统进行译文自动后编辑，与基于短语的统计后编辑模型[5]相比，该方法极大的提高了译文后编译的效果。为了考虑翻译的上下文信息，Pecina 等人[12]在双向递归神经网络编码器-解码器模型的基础上，为源语言句子和机器译文分别构建独立的编码器，并加权融合这两个编码器所得到的上下文向量 (word embedding) 作为解码信息，该方法与多源神经机器翻译方法[13]非常相似，最大的区别在于多源信息使用的是待翻译的源语言句子和其待编辑的机器译文。Grundkiewicz 等人[14]提出将单语与双语神经机器翻译模型输出进行对数线性联合以提高译文后编辑的质量。

为了有效控制译文后编辑过程中的过度修正问题，本文在 Grundkiewicz 等人[14]的工作基础上，对机器译文中需要少量编辑次数的句子构建了专门的神经网络自动后编辑模型，并利用译文质量估计方法从后编辑模型输出的译文 *n*-best 列表中挑选最优后编辑译文。

2 译文后编辑语料的统计与分析

根据 TER 的定义[7]，译文修改所使用到的编辑操作被分为词语插入(insertion)，删除(deletion)，替换(substitution)，语块移位(shift of word chunk)四种类型。在实际的译文自动后编辑过程中，大多数句子的修正都是通过复合使用这四种编辑操作来完成，同时也有部分句子仅需使用一次或多次单一类型的编辑操作即可完成修正。

在自动后编辑中，过度修正是影响后编辑译文质量的主要原因之一。为了探究后编辑的规律，并预测出 WMT16 APE 测试集中 2000 条机器译文所需的后编辑形式，我们从“编辑次数”和“编辑类型”两个角度对 WMT 官方发布的训练集语料进行了如下的统计分析：

利用 TER 脚本程序，我们对训练集中未经过处理的 23000 条机器译文所需的编辑次数进行了统计，结果如图 1 所示，将机器译文修正成正确译文平均所需的编辑次数为 4，其中有 58.03%的句子所需的编辑次数不超过 4 次。一个特别的发现是：有 20.47%的机器译文修正成正确译文所需的编辑次数不超过 1 次。这意味着在该语料中可能存在一定规模的数量的句子在修正过程中仅需单一类型的编辑操作。

为了进一步对机器译文中需要复合编辑操作和需要单一编辑操作的句子的特点进行研究，我们统计了后编辑过程中仅需要使用单一类型编辑操作的句子，图 1 的结果表明在 4514 条仅需单一类型编辑操作的机器译文句子中，有超过 80%的句子所需的编辑操作次数小于或等于 2。

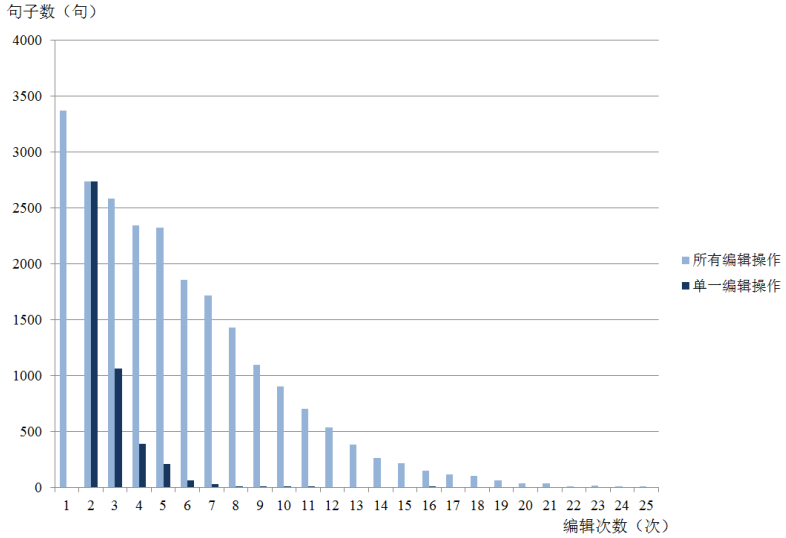

图 1　WMT 训练集中需要复合编辑操作与单一编辑操作的机器译文所需编辑次数的分布情况

3　模型

在训练集上编辑操作的分布情况表明，存在很多机器译文只需要少量的复合编辑操作就可以将其转换为正确译文，而且，还有一些机器译文只需要少量的单一编辑操作就可以将其转换为正确译文；因此，我们预测这种现象也存在于测试集上，并且，我们分别对这两种情况进行建模。

3.1 神经翻译模型

在神经机器翻译中，Sutskever 等人[15]提出基于长短时记忆网络的编码器-解码器模型，该模型工作时，编码器将源语言句子编码为一个上下文向量，而解码器根据该向量和已经生成的目标语言单词挑选出概率最大的后续单词作为译文输出。Bahdanau 等人[16]提出了双向递归神经网络编码器对上述模型进行了扩展，并引入了注意力机制[17]，在该模型中，源语言上下文向量不再是由编码结果中的单一状态决定，而是根据注意力机制对源语言句子中的单词赋予了不同的权重。

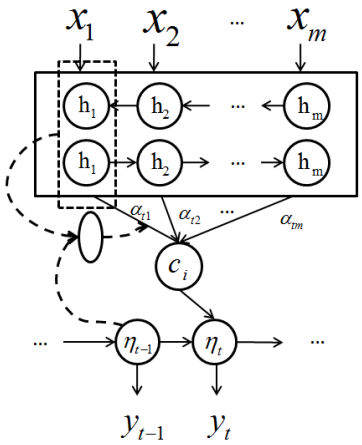

图 2　带有注意力机制的双向递归神经网络编码器-解码器模型

带有注意力机制的双向递归神经网络编码器-解码器模型由两个不同方向的递归神经网络层组成，其网络结构如图 1 所示，对于需要读入的句子序列 x ($x = x_1, x_2, \ldots, x_m$)，每一时刻分别从源语言句子的

正向和逆向依次读入单词 x_i，递归神经网络的隐状态由一个非线性激活函数进行更新，该激活函数由一个长短期记忆单元与一个 Sigmoid 激活函数组成，如公式(1)所示，其中，h_{t-1} 表示递归神经网络在 t-1 时刻的隐状态，x_i 表示在 t 时刻读入的单词，t 时刻的隐状态 h_t 由 h_{t-1} 与 x_t 联合确定。

$$h_t = f(h_{t-1}, x_t) \tag{1}$$

读入待翻译的源语言句子后，通过递归神经网络的隐状态可以得到包含整个句子信息的上下文向量（图 2 中由 c_i 表示）。解码部分由单层递归神经网络构成，其工作是根据编码过程得到的上下文向量以及当前的隐状态预测出最有可能的待生成词。解码过程 t 时刻的隐状态可通过公式(2)和公式(3)计算：

$$P(y_t \mid y_1, ...y_{t-1}, \mathrm{x}) = f(\eta_t, y_{t-1}, c_t) \tag{2}$$

$$\eta_t = f(\eta_{t-1}, y_{t-1}, c_t) \tag{3}$$

上下文向量 c_t 可由公式(4)计算：

$$c_t = \sum_{i=1}^{m} \alpha_{ti} h_i \tag{4}$$

其中 a_{ti} 表示每个隐状态 h_i 的权重，其计算方法如下：

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{m} \exp(e_{tj})} \tag{5}$$

$$e_{ti} = a(\eta_{t-1}, h_i) \tag{6}$$

$$a(\eta_{t-1}, h_i) = \upsilon_a^T \tanh(W_a \eta_{t-1} + U_a h_i) \tag{7}$$

其中，e_{ti} 表示编码过程中输入句子中位置 i 与解码过程中输出句子中位置 t 的对齐得分，这个得分由公式(7)给出，在整个句子的解码过程中，该得分共计算了 $m \times n$ 次。

3.2 神经网络自动后编辑模型

在 Grundkiewicz 等人[14]的工作基础上，我们使用带有注意力机制的双向递归神经网络编码器-解码器模型分别建立了一个单语后编辑模型和一个双语翻译模型，其中单语后编辑模型使用机器译文与人工后编辑译文组建的平行语料进行训练，双语翻译模型则使用源语言句子与人工后编辑译文组建的平行语料进行训练，并通过使用对数线性联合的方式将两者进行融合。该模型通过 WMT 官方提供的训练集对其进行训练，其在实验中被作为对比的基准系统 ($\mathrm{NPE_{BASELINE}}$)。

为了减少过度修正，根据在训练集上后编辑操作的分布情况，我们从官方训练集中抽取出的所需复合编辑次数不超过 4 次和单一编辑操作次数不超过 2 次的机器译文句子分别构成训练集，通过对基准模型进行微调的方式训练得到了用于提供少量复合编辑操作的神经网络后编辑系统 ($\mathrm{NPE_{MINOR}}$)，以及提供少量单一后编辑操作的神经网络后编辑系统($\mathrm{NPE_{SINGLE}}$)。

3.3 机器翻译质量估计

由于机器译文很难事先通过自动分类的方法确定其编辑类型，再采用具体的后编辑模型进行处理，因此我们汇集不同神经网络后编辑模型对同一待后编辑的机器译文的编辑结果组成 n-best 列表，利用机器译文质量估计方法从后编辑 n-best 列表中筛选挑选最优后编辑译文。

机器翻译质量估计 (Quality Estimation, QE) 是机器翻译领域新兴起的一个研究方向，与传统的机器翻译质量评价方法[18-20]不同，质量估计方法评估机器译文的质量无需使用人工参考译文。句子级别的译文质量估计一般被看作是机器学习中的回归问题，其任务目标是利用从源语言句子和机器译文中抽取描述译文质量的特征，包括反映翻译复杂度、流利度和忠实度的特征等，用于回归求取译文质量[21]。

神经网络译文质量估计方法[22-23]被用来对后编辑机器译文进行排序打分，从后编辑 *n*-best 列表中挑选最优后编辑译文。

在完成译文质量估计模型的训练之后，我们对译文质量估计打分的精确性进行了评估。通过在不同规模的 *n*-best 句子集合上使用该得分进行排序测试发现：(1) 当参与排序的译文质量差异较大时，译文质量估计方法可以有效的对句子质量的优劣进行区分，但是当译文质量之间的差异较小时，其排序准确性将会出现一定程度的下降；(2) 由于 *n*-best 中的译文质量整体的不同，造成译文质量估计的得分差异也随之变化。

为了有效发挥质量估计的作用，降低误估带来的影响，参照人工对机器译文质量的“五分制”评价标准（5 分表示最高质量得分），我们采用了一个译文质量分层的方法来完成对 *n*-best 列表中优质候选译文的筛选，如图 4 所示，对于待后编辑的机器译文，首先使用一个基于译文质量估计的模型对它进行质量打分，在得到质量得分之后，对各个自动后编辑模型在该质量验证集上的 TER 得分进行排序，并根据排序结果从 *n*-best 列表中选择最优候选译文作为后编辑输出。此外，当各模型在验证集上的表现差异不显著时，我们则通过引入统计语言模型 SRILM 打分的方式对评估结果进行进一步的强化。

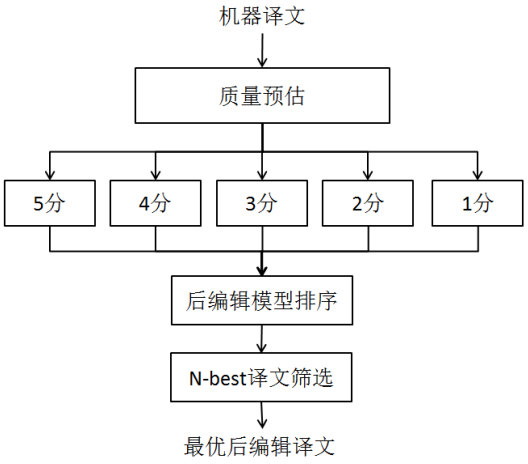

图 4　基于 QE 的质量分层排序方法流程

4　实验

为了测试本文所提出方法的效果，我们在 WMT16 APE 子任务的测试集上进行了实验验证。

4.1 实验设置

实验使用的语料包括 WMT16 APE 子任务官方发布的语料与 Grundkiewicz 等[14]公开发布的自动后编辑语料。语料均由源语言句子，机器译文和人工后编辑译文的三平行句子组成，语料的具体统计信息如表 1 所示。

表 1　实验使用数据集的规模统计

语料名称	句子数（平行语料）	机器译文平均词数	TER
WMT16 训练集	12,000	17.89	26.22
WMT16 开发集	1,000	19.75	24.81
WMT16 测试集	2,000	17.41	24.76
人工后编辑语料 500K	531,839	20.92	25.28
人工后编辑语料 4M	4,335,715	15.86	36.63

由于 WMT 的训练集语料规模较小，一共只有 15000 句三平行语料，使用它训练机器翻译后编辑系统将导致数据稀疏，Grundkiewicz 等[14]在 Common Crawl 语料德语子语料上通过交叉熵过滤以及双向翻译构建了译文自动后编辑语料对官方语料进行扩充，并将扩充的语料公开发布，实验中融合了这两个语料作为训练集，用以训练神经网络自动后编辑系统。

根据本文 3.2 节的模型设计，我们从上述数据集中抽取出用于训练 $NPE_{BASELINE}$，NPE_{MINOR} 和 NPE_{SINGLE} 的子训练集构成如表 2 所示：

表 2　各子模型训练集的规模统计

系统名称	训练集规模（平行语料）	机器译文平均词数	TER
$NPE_{BASELINE}$	4,867,554	16.31	35.39
NPE_{MINOR}	2,235,959	12.73	15.98
NPE_{SINGLE}	1,134,963	11.87	8.79

为了使词汇表更具有覆盖性，源语言英语句子和目标语言德语句子均进行了标记化 (tokenizer) 和大小写 (truecase) 处理。为了提高系统性能，句子均被转换为子词序列 (subword units) [24]，在译文自动后编辑中，使用子词单元作为基本单元进行处理。

我们使用 Nematus[25]对带有注意力机制的双向递归神经网络模型的编码器进行训练，模型参数中批量处理的句子数量设置为 80，词表规模设置为 40000，句子的最大长度限定为 50，词向量的维数设置为 500，神经网络隐含单元个数设置为 1024，使用 Adadelta 算法[23]进行模型的参数优化。与 Nematus 方法相比，基于 C++/CUDA 的 AmuNMT 方法[14]在 CPU 上进行解码的速度更快，因此在模型的解码部分，AmuNMT 方法被利用对待编辑的机器译文进行解码，其中 beam 设置为 12，并对句子长度进行了标准化。

4.2 实验结果

表 3　在 WMT16 APE 子任务测试集上不同系统的性能

系统名称	TER	BLEU
Baseline1(Raw MToutput)	24.76	62.11
Baseline2(Moses PBAPE)	24.64	63.47
NPE_{MTPE}	24.28	64.35
$NPE_{BASELINE}$	**23.78**	**64.97**
$NPE_{BASELINE}$ + NPE_{MINOR} + $RESCORE_{QE}$	23.08	66.73
$NPE_{BASELINE}$ + NPE_{MINOR} + NPE_{SINGLE} + $RESCORE_{QE}$	**22.33**	**67.34**
$NPE_{BASELINE}$ + NPE_{MINOR} + NPE_{SINGLE} + $RESCORE_{SRILM}$	23.15	66.18

表 4　联合模型中各子模型在 QE 重排序模型中的排名情况（取排名前两位的子模型）

质量预估打分	QE 得分区间	最优子模型	次优子模型	最优与次优模型的 TER 差值	是否使用 SRILM 排序
5	<=13.3	NPE_{MINOR}	NPE_{SINGLE}	0.60	否
4	(13.3,17.35]	NPE_{SINGLE}	NPE_{MINOR}	0.22	是
3	(17.35,23.45]	NPE_{MINOR}	NPE_{SINGLE}	0.46	否
2	(23.45,25.45]	NPE_{MINOR}	$NPE_{BASELINE}$	0.19	是
1	>25.45	NPE_{MINOR}	$NPE_{BASELINE}$	0.58	否

首先，将训练得到的 $NPE_{BASELINE}$ 系统与 WMT16 APE 子任务官方发布的 baseline 进行对比，从表 3 中可以看到，基准系统 $NPE_{BASELINE}$ 比官方发布的 Baseline1(Raw MT output) (未做后编辑的原始机器

译文)与 Baseline2(Moses PBAPE) (基于短语的统计后编辑模型)以及 NPE_{MTPE}(单语神经自动后编辑模型)在 TER 和 BLEU 得分上均有明显提升。

其次，将处理少量复合后编辑操作的神经网络后编辑模型 NPE_{MINOR} 与基准系统 $NPE_{BASELINE}$ 进行融合，融合的后编辑系统 $NPE_{BASELINE}$ + NPE_{MINOR} 比基准系统 $NPE_{BASELINE}$ 翻译编辑率 TER 降低了 0.7，BLEU 值提高了 1.76；最后，将处理少量单一后编辑操作的神经网络后编辑模型 NPE_{SINGLE} 与上述复合模型进一步融合，模型 $NPE_{BASELINE}$ + NPE_{MINOR} + NPE_{SINGLE} 进一步提高了译文的质量，相比基准系统 $NPE_{BASELINE}$，翻译编辑率 TER 降低了 1.45，BLEU 值提高了 2.37，统计显著性检验结果表明，在置信区间 p=0.05 情况下，该提升是统计显著的。

为了评价基于 QE 的排序模型，表 4 给出了联合模型中各子模型在 QE 分层排序中的性能，结果表明基于 QE 的排序模型在 BLEU 和 TER 两个评价指标上均取得了更优的结果。

4.3 实验分析

为了揭示基于译文质量估计的神经网络后编辑系统提高译文质量的原因，实验进一步比较了模型 $NPE_{BASELINE}$ + NPE_{MINOR} + NPE_{SINGLE} 与基准系统 $NPE_{BASELINE}$ 在测试集后编辑操作次数上的分布。

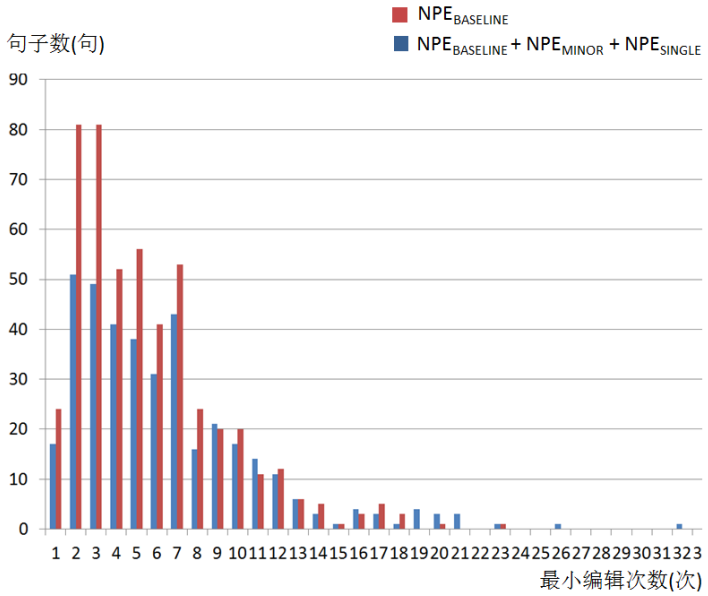

图 5　在 WMT16 APE 子任务测试集上基准模型与联合模型的编辑次数分布

实验发现基准模型输出的后编辑译文中，有 500 条后编辑译文相比原始机器译文在 TER 得分上出现了上升 （表明编辑后译文质量出现了下降），我们对这 500 条出现过度修正的机器译文所需的最小编辑次数分布进行了统计。结果表明，在出现过度修正的机器译文中，有超过 58%的句子所需的编辑次数不超过于 4 次，反映出过度修正情况主要出现在对需要较少后编辑的机器译文的修正中。而在联合模型的输出译文中，出现过度修正情况的句子一共有 372 句。这说明相比使用单一的神经网络后编辑模型而言，联合模型在减少过度修正的问题上有比较明显的效果。为了更加直观的揭示联合模型在减少过度修正方面的作用，我们对基准系统 $NPE_{BASELINE}$ 与联合模型 $NPE_{BASELINE}$ + NPE_{MINOR} + NPE_{SINGLE} 在后编辑过程中出现过度修正情况的机器译文所需的最小编辑次数进行了统计，从图 5 中可以发现，联合模型在对需要少量(<=4)编辑次数的机器译文进行修正时，出现过度修正的频率明显少于基准模型。

5　结论

为了减少译文过度修正，在数据分析的基础上，本文提出了建立不同的神经网络后编辑模型，包括处理少量复合后编辑操作的模型和处理少量单一后编辑操作的模型，并通过最新的译文质量估计方法对融合后的联合模型进行分级筛选。实验结果表明该方法能有效地提高后编辑的译文质量。该方法的一个

不足之处在于它仅针对 WMT APE 任务中英语-德语翻译方向的子任务进行数据分析和建模，在其他语言对的翻译任务上，后编辑操作次数的分布可能与该任务有差异，导致该模型难以得到直接应用。因此，在以后的工作中，我们将对模型的泛化性进行进一步的研究。

参考文献

[1] 刘群. 基于句法的统计机器翻译模型与方法. 中文信息学报, 2011, 25(6): 63-72.

[2] 冯志伟. 机器翻译研究. 中国对外翻译出版公司, 2004.

[3] 宗成庆. 统计自然语言处理. 清华大学出版社, 2008.

[4] Knight K, Chander I. Automated postediting of documents // AAAI. 1994: 779-784.

[5] Simard M, Goutte C, Isabelle P. Statistical phrase-based post-editing. 2007: 508-515.

[6] Béchara H, Ma Y, van Genabith J. Statistical post-editing for a statistical MT system // MT Summit. 2011, 13: 308-315.

[7] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation // Proceedings of association for machine translation in the Americas. 2006: 223-231.

[8] Chatterjee R, Weller M, Negri M, et al. Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing // Proceedings of the ACL, 2015: 156-161

[9] Ondˇrej Bojar, Rajen Chatterjee, et al. Findings of the 2015 Workshop on Statistical Machine Translation // Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015: 1–46.

[10] Bojar O, Chatterjee R, Federmann C, et al. Findings of the 2016 conference on machine translation // Proceedings of the First Conference on Machine Translation (WMT). 2016, 2: 131-198.

[11] Pal S, Naskar S K, Vela M, et al. A neural network based approach to automatic post-editing // Proceedings of the ACL. 2016: 281-286.

[12] Libovicky J, Helcl J, Tlusty M, et al. CUNI system for wmt16 automatic post-editing and multimodal translation tasks. arXiv preprint arXiv: 1606.07481, 2016.

[13] Zoph B, Knight K. Multi-source neural translation. arXiv preprint arXiv: 1601.00710, 2016.

[14] Junczys-Dowmunt M, Grundkiewicz R. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. arXiv preprint arXiv: 1605.04800, 2016.

[15] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. 2014: 3104-3112.

[16] Cho K, Van Merriënboer B, Gulcehre C et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1724–1734.

[17] Bahdanau D, Cho K, Bengio Y et al. Neural machine translation by jointly learning to align and translate // Proceedings of the International Conference on Learning Representations, 2015,San Diego, CA.

[18] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, 2002: 311–318.

[19] Li M, Wang M, Li H, et al. Modeling monolingual character alignment for automatic evaluation of Chinese translation. ACM Transactions on Asian and Low-Resource Language Information Processing, 2016, 15(3): 1–18.

[20] 张丽林, 李茂西, 肖文艳, 等. 机器翻译自动评价中领域知识复述抽取研究. 北京大学学报 (自然科学版), 2017, 53(2): 230-238.

[21] Specia L, Shah K, De Souza J G C, et al. QuEst - A translation quality estimation framework // Proceedings of the ACL (Conference System Demonstrations). 2013: 79-84.

[22] Shah K, Bougares F, Barrault L, et al. SHEF-LIUM-NN: Sentence level Quality Estimation with Neural Network Features // WMT. 2016: 838-842.

[23] 陈志明，李茂西，王明文．基于神经网络特征的句子级别译文质量估计．计算机研究与发展，2017，54(8): 1804-1812.

[24] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv: 1508.07909, 2015.

[25] Sennrich R, Firat O, Cho K, et al. Nematus: a toolkit for neural machine translation. arXiv preprint arXiv: 1703.04357, 2017.

[26] Zeiler M D. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv: 1212.5701, 2012.